Uppsala University at SemEval-2023 Task12: Zero-shot Sentiment Classification for Nigerian Pidgin Tweets

Annika Kniele and Meriem Beloucif
Department of Linguistics and Philology
Uppsala University
annika.kniele.4937@student.uu.se

Abstract

While sentiment classification has been considered a practically solved task for high-resource languages such as English, the scarcity of data for many languages still makes it a challenging task. The AfriSenti-SemEval shared task aims to classify sentiment on Twitter data for 14 low-resource African languages. In our participation, we focus on Nigerian Pidgin as the target language. We have investigated the effect of English monolingual and multilingual pre-trained models on the sentiment classification task for Nigerian Pidgin. Our setup includes zero-shot models (using English, Igbo and Hausa data) and a Nigerian Pidgin fine-tuned model. Our results show that English fine-tuned models perform slightly better than models fine-tuned on other Nigerian languages, which could be explained by the lexical and structural closeness between Nigerian Pidgin and English. The best results were reported on the monolingual Nigerian Pidgin data. The model pre-trained on English and fine-tuned on Nigerian Pidgin was submitted to Task A Track 4 of the AfriSenti-SemEval Shared Task 12, and scored 25 out of 32 in the ranking.

1 Introduction

Nigerian Pidgin (*pcm*) is an English-based Creole language spoken in Nigeria by ca. 40 million native speakers and ca. 80 million second-language speakers (Muhammad et al., 2022). Table 1 shows an example of a Nigerian Pidgin tweet:

pcm	E don tay wey I don dey crush on this fine woman. . .
en	*I have had a crush on the beautiful woman for a while. . .*

Table 1: Example of a tweet in Nigerian Pidgin (Muhammad et al., 2022). We note that there are few words that an English speaker can easily understand and distinguish.

Despite having a considerable number of speakers, it is considered a *low-resource* language. Due to lexical and structural similarities between English (*en*) and Nigerian Pidgin (Muhammad et al., 2022), one might assume that English could be a good candidate language to be used in zero-shot experiments for Nigerian Pidgin sentiment classification. Since English is a high-resource language, this could go some way towards alleviating the resource scarcity problem for Nigerian Pidgin.

Besides Nigerian Pidgin, two of the most widely spoken languages in Nigeria are Igbo (*ig*) and Hausa (*ha*). The Igbo language is part of the Benue-Congo group of the Niger-Congo languages, and Hausa is one of the Chadic (Afroasiatic) languages. This means that while the two languages are geographically close to Nigerian Pidgin, no close linguistic ties exist between them and Nigerian Pidgin (Muhammad et al., 2022). However, code-switching between the two languages and Nigerian Pidgin, or English, is not uncommon. As an example, consider Table 2, which gives examples of tweets in Igbo and Hausa, both of which are code-switched with Nigerian Pidgin:

ig	akowaro ya ofuma nne kai daalu nwanne mmadu we go dey alright las las
en	*they told it well my fellow sister well done at the end we will be all right*
ha	Aunt rahma i luv u wallah irin totally dinnan
en	*Aunty rahma I swear I love you very much*

Table 2: Example of tweets in Igbo and Hausa (Muhammad et al., 2022, p. 592).

In this paper, we investigate how using different languages for pre-training and fine-tuning affects the performance of zero-shot sentiment classification for Twitter data in Nigerian Pidgin. We also compare different models to the standard monolingual sentiment classification task. The best monolingual result in this paper was submitted to Task

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1491–1497

A Track 4 of the AfriSenti-SemEval Shared Task 12 (Muhammad et al., 2023b).[1]

2 Background

In recent years, with advances in the field of deep learning, the performance of *Natural Language Processing* (NLP) applications has been improving steadily (Hedderich et al., 2020). However, models that use these architectures require a lot of data to learn from. For languages deemed low-resource, this amount of data is often not available. To address this issue, several approaches using additional resources have been proposed in the literature, such as transfer learning and data augmentation (Hedderich et al., 2020).

Sentiment classification is a popular NLP task which aims to determine the sentiment expressed in a text (e.g. positive, negative, neutral). There has been a lot of interest in sentiment classification tasks for low-resource languages in recent years, with work done on a diverse range of languages such as Persian (Ghasemi et al., 2022), Tigrinya (Fesseha et al., 2021), Hindi (Kumar and Albuquerque, 2021) and Uzbek (Kuriyozov et al., 2019).

While there has also been some research in this area for the low-resource language Nigerian Pidgin (Oyewusi et al. 2020, Adamu et al. 2021), the number of studies, especially when it comes to deep learning approaches, is limited. One example of such an approach is Muhammad et al. (2022), who worked on sentiment classification for the four languages with the highest number of speakers in Nigeria: Hausa, Igbo, Nigerian Pidgin and Yorùbá. They first collected and annotated a sentiment Twitter corpus. Subsequently, they carried out a number of experiments on sentiment classification using several multilingual pre-trained language models, namely mBERT, a multilingual version of BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), RemBERT (Chung et al., 2021), AfriBERTA (Ogueji et al., 2021) and mDeBERTaV3 (He et al., 2021). They experimented with zero-shot fine-tuning on an English Twitter dataset (English SemEval 2017 Task 4, Rosenthal et al. 2017), and monolingual as well as multilingual supervised fine-tuning using only the Nigerian languages specified above. For the zero-shot experiment, the highest F1-score for Nigerian Pidgin was achieved by the mDeBERTaV3 model (He et al., 2021), the only model used in Muhammad et al.'s (2022) study not pre-trained on any Nigerian languages. Herein, we investigate whether exclusively using English for both pre-training and fine-tuning has some advantages over using multiple languages, including other African ones.

[1]The code for this project can be found at github.com/akniele/sentiment_classification_Nigerian_Pidgin.

3 Data

3.1 Datasets

The data used in this paper is taken from two different shared tasks: Firstly, the data for Nigerian Pidgin, Igbo and Hausa is provided by Muhammad et al. (2023a). Each tweet in the dataset is labeled as either "positive", "negative", or "neutral". For Igbo and Hausa, we concatenate their respective train and dev sets, to be able to use them together for fine-tuning.

Secondly, the labeled data for sentiment classification for English is taken from the development set provided for SemEval-2017 Task 4 (Rosenthal et al., 2017), and split into train and dev (80:20). Just like the datasets for the Nigerian languages, this is also a Twitter dataset, with each tweet labeled as "positive", "negative", or "neutral". Table 3 gives an overview of the number of tweets for each language setup. We can see that the Igbo-Hausa dataset is the largest, followed by the English dataset. The Nigerian Pidgin dataset is quite small in comparison to the other two:

	train	**dev**	**test**
pcm	5121	1281	4154
en	22753	5689	-
ig+ha	24364	4518	-

Table 3: Number of tweets

3.1.1 Possible issues with the datasets

Muhammad et al. (2023a), who collected the datasets for the AfriSenti shared task, note that the distribution of classes in some of their datasets is quite imbalanced. Figure 1 shows the distribution for the datasets (train and dev) used in this work. In the Nigerian Pidgin dataset, there are 2,255 positive, 4,054 negatives, and only 93 neutral tweets. This majorly influences how well a model fine-tuned on this data will predict neutral tweets, seeing as in imbalanced datasets, learning the minority classes can be challenging (Shi et al., 2022). Furthermore, in the English dataset, as well as in the Igbo-Hausa dataset, "neutral" is the most

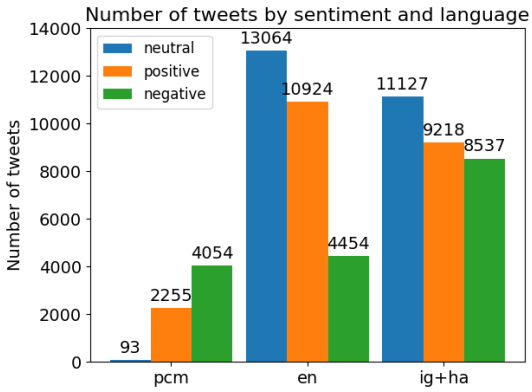

Figure 1: Class distribution for the different languages

common label. When fine-tuning the model on the English or Igbo-Hausa dataset, and then testing it on Nigerian Pidgin, the model might overpredict the category "neutral", which could equally have a detrimental effect on its performance. Some options for alleviating this would be undersampling the bigger classes (and thereby creating a more balanced dataset), or using data augmentation (Feng et al., 2021). However, these options are not explored further in this paper.

3.2 Preprocessing

In part following Muhammad et al. (2022), duplicates, URLs, mentions, as well as trailing and other redundant white spaces are deleted. Punctuation is also removed and the data is lowercased. Furthermore, as there appeared to be a considerable number of emojis in the data, these are removed using the emoji library available for Python[2]. After this step, there remained some Unicode symbols, such as musical notes and hearts, as well as a small number of Chinese characters. All those were manually collected into a list and subsequently automatically deleted from the data. Finally, there was some Arabic writing in the data, which was omitted using the PyArabic library (Zerrouki, 2010).

4 Experimental setup

4.1 Models

In this paper, two BERT (Devlin et al., 2019) models are used: Firstly, the monolingual *bert-base-uncased* model (Devlin et al., 2019), which has exclusively been pre-trained on English data, and secondly, the multilingual *bert-base-multilingual-uncased* model (Devlin et al., 2019), which has been pre-trained on 102 languages, excluding Nigerian Pidgin, Igbo and Hausa. These models have been chosen as they have a similar architecture, and this could facilitate comparing the results of the different setups. Both the models used in this paper are available through the Huggingface *Transformers* library [3] [4], an open-source library for transformer models (Wolf et al., 2020).

4.2 Implementation

The experimental setup used by Muhammad et al. (2022) was partially replicated. They used the mBERT-base-cased model and fine-tuned it on English data for sentiment classification, before testing it on Nigerian Pidgin data for the same task. However, since parts of the data obtained from the shared task were lowercased from the outset, using the cased versions of the BERT models, as was done by Muhammad et al. (2022), would not have been a good choice for this paper, seeing as the case information was already lost for part of the data. Instead, in this paper, the rest of the data are also lowercased, and the uncased versions of the BERT model are used.

Table 4 shows an overview of the different setups. The first column contains a shorthand name for each setup. The second column shows which pre-trained model was used, the third column the fine-tuning language, and the fourth column the test language, which was Nigerian Pidgin in all cases. Note that the two setups for which the model is fine-tuned and tested on Nigerian Pidgin are not zero-shot. They do, however, provide a baseline to which the zero-shot setups can be compared, as well as allow us to compare the effect of using a monolingual or a multilingual pre-trained model.

shorthand	pre-train	train & val	test
mono-en	BERT	en	pcm
mono-pcm	BERT	pcm	pcm
mono-igha	BERT	ig & ha	pcm
multi-en	mBERT	en	pcm
multi-pcm	mBERT	pcm	pcm
multi-igha	mBERT	ig& ha	pcm

Table 4: Experimental setups for this paper

Furthermore, the hyperparameters used for the model submitted to the shared task are the same ones as those used by Muhammad et al. (2022,

[2] https://pypi.org/paper/emoji/

[3] https://huggingface.co/bert-base-uncased

[4] https://huggingface.co/bert-base-multilingual-uncased

p. 602), namely a batch size of 32, a maximum sequence length of 128, 20 epochs, and a learning rate of 5e-5. All other models are trained with the same hyperparameters, but only for 4 epochs to reduce overfitting. We also fixed the dropout at 0.5.

Finally, to evaluate the results, the weighted F1-score is used as the main evaluation metric, implemented using the Python machine learning library *sklearn*[5].

5 Results

In the interest of reproducibility, the results reported in this section and Section 6 are averages obtained over four runs, with the PyTorch random seed set to 2,3,4 and 5, respectively. The test accuracy of the model submitted to the shared task is **68.8%**.

Table 5 shows the test accuracy for the different setups:

	mono	**multi**
pcm	68.9	66.8
en	32.5	35.1
ig+ha	32.3	31.6

Table 5: Test accuracy (%)

It is not so surprising that fine-tuning on Nigerian Pidgin, and subsequently testing on the same language, gave the highest test accuracy, with 68.9% for the mono-pcm setup, and 66.8% for multi-pcm. When looking at the zero-shot scenarios, it can be seen that fine-tuning on English gives a slightly higher test accuracy, with 32.5% and 35.1% for mono-en and multi-en, respectively. Fine-tuning on Igbo and Hausa yielded a lower test accuracy, with 32.3% for mono-igha, and 31.6% for multi-igha.

The English scenario yielding slightly better results than the Igbo-Hausa scenario could be explained with respect to the close linguistic proximity between English and Nigerian Pidgin, which could have had a larger positive impact than the close geographical proximity between Igbo, Hausa, and Nigerian Pidgin. However, it is worth noting that there was a high variability in the scores these averages are based on, meaning that this effect might be due to the specific random seeds chosen.

Furthermore, when it comes to the choice of pre-trained model, it can be noted that when fine-tuning on Nigerian Pidgin or Igbo and Hausa, using BERT-base-uncased seems slightly preferable to using mBERT-base-uncased, while for fine-tuning on English, the test accuracies suggest a slight advantage in using mBERT-base-uncased. In general, due to the high variability of the individual scores, any conclusions should be tentative.

Table 6 shows the F1-scores for the different setups. As with the test accuracies, the F1-scores also suggest that when fine-tuning on Nigerian Pidgin or Igbo and Hausa, using BERT-base-uncased for this task is preferable to using the multilingual variant, while for fine-tuning on English mBERT-base-uncased seems more suitable. Furthermore, fine-tuning on Nigerian Pidgin again yielded the best results, with 65.0% (**64.7%** in the shared task) and 62.5% F1-score, respectively. In the English zero-shot scenario, the models achieved an F1-score of 36.9% and 39.6%, respectively. For Igbo and Hausa, the scores are 36.3% and 33.3%. As for the test accuracies, one could also conclude here that the lexical similarity between English and Nigerian Pidgin might explain why this scenario received slightly better results than Igbo and Hausa, or it could be due to the arbitrarily chosen random seeds.

	mono	**multi**
pcm	65.0	62.5
en	36.9	39.6
ig+ha	36.3	33.3

Table 6: F1-scores (%)

6 Analysis

As the classes in the datasets are quite imbalanced, we might ask ourselves how well the models predict the individual classes. Due to space considerations, only the *recall* (i.e. the percentage of tweets of each class that was correctly classified), is considered in this section. Table 7 shows recall when testing the models fine-tuned on Nigerian Pidgin. As we can see, both the mono-pcm and multi-pcm setups recall more than 80% of the negative tweets. This is not surprising, seeing as the negative tweets are by far the biggest class in the dataset. Positive tweets also receive a fairly high percentage score, with 65.5% for mono-pcm and 57.0% for multi-pcm. On the other hand, almost none of the neutral tweets were correctly identified by the models. This might be because only about 1.5% of the train/dev data for Nigerian Pidgin is neutral tweets, whereas the

[5] https://scikit-learn.org/stable/

test data contains about 10% neutral tweets.

	mono	multi
negative	83.6	85.0
neutral	0.17	0.0
positive	65.5	57.0

Table 7: Recall when fine-tuning on *pcm* (%)

Table 8 shows the recall when fine-tuning on English. In this scenario, only 23.2% and 29.2% of the negative tweets were recalled by the models. This is considerably lower than when fine-tuning on Nigerian Pidgin and could be due to the negative class being by far the smallest one in the English dataset. When it comes to positive tweets, both models recalled 30.8% of them. This is considerably lower than when fine-tuning on Nigerian Pidgin. Finally, the models recalled 88.6% and 80.6% of the neutral tweets, respectively. This is not surprising, seeing as neutral tweets are the biggest class in the English train and development set, comprising about 45.9% of the tweets. This shows that adding the English data helped to add an equilibrium when it comes to the neutral tweets in contrast to the Nigerian Pidgin data.

Table 9 shows the recall when fine-tuning on Igbo and Hausa. When it comes to negative tweets, we can see that mono-igha performed better than its monolingual counterpart which was fine-tuned on English (27.1% and 23.2%, respectively). Conversely, multi-igha achieved a much lower recall than the multilingual model fine-tuned on English (20.7% and 29.2%, respectively). As mentioned above, the high variability of the individual scores makes it difficult to draw conclusions about why this might be.

The recall of positive tweets was quite a bit lower for mono-igha than for mono-en, with 24.4% recall compared to 30.8%. Using the mBERT-base-uncased, however, the Igbo and Hausa fine-tune model outperformed the one fine-tuned on English with 34.3% recall (compared to 30.8%). For the neutral tweets, both the monolingual and the multilingual models recalled between 81% and 86% of the tweets. One reason why they performed so well on the neutral tweets could be that just like with the English dataset, "neutral" is the largest class in the Igbo-Hausa dataset, and therefore the models might be over-predicting this class.

	mono	multi
negative	23.2	29.2
neutral	88.6	80.6
positive	30.8	30.8

Table 8: Recall when fine-tuning on *en* (%)

	mono	multi
negative	27.1	20.7
neutral	85.8	81.6
positive	24.4	34.3

Table 9: Recall when fine-tuning on *ig* and *ha* (%)

7 Conclusion

This paper has investigated different zero-shot scenarios for sentiment classification in the low-resource language Nigerian Pidgin. It has been shown that no clear conclusions can be drawn regarding whether it is preferable to use a monolingual or a multilingual pre-trained model in the setups investigated in this paper. Furthermore, while using Nigerian Pidgin data for fine-tuning yields far better results than using other languages, a zero-shot approach using English data yields only slightly better results than using data in Igbo and Hausa. This difference could either be due to linguistic proximity between English and Nigerian Pidgin or might have come about by chance due to the large variability in scores every time the models are run. Whichever the case, it highlights the importance of having access to monolingual data as opposed to solely relying on data from related languages. The results also emphasize the importance of having balanced training data, at least in a low-resource setting. In how far the results presented in this report were influenced by the classes in the dataset being imbalanced could be explored further by experimenting more with balancing out the classes, either by downsampling or by adding more data, already existing or synthetically created. Another further research direction could investigate the effect of using both Nigerian Pidgin and English data for fine-tuning.

Acknowledgements

The computations in this paper were run on the Uppsala Multidisciplinary Center for Advanced Computational Science, through project UPPMAX 2020/2-2.

References

Hassan Adamu, Syaheerah Lebai Lutfi, Nurul Hashimah Ahamed Hassain Malim, Rohail Hassan, Assunta Di Vaio, and Ahmad Sufril Azlan Mohamed. 2021. Framing Twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning. *Sustainability*, 13(6):3497.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Awet Fesseha, Shengwu Xiong, Eshete Derb Emiru, Moussa Diallo, and Abdelghani Dahou. 2021. Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Information*, 12(2):52.

Rouzbeh Ghasemi, Seyed Arad Ashrafi Asli, and Saeedeh Momtazi. 2022. Deep Persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal of Information Science*, 48(4):449–462.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309v3*.

Akshi Kumar and Victor Hugo C. Albuquerque. 2021. Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor Indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–13.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2019. Deep learning vs. classic models on a new Uzbek sentiment analysis dataset. *Proceedings of the Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 258–262.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Shehu Bello Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahuddeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic enrichment of Nigerian Pidgin English for contextual sentiment classification. *arXiv preprint arXiv:2003.12450v1*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Yiwen Shi, Taha ValizadehAslani, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou

Liang. 2022. Improving imbalanced learning by pre-finetuning with data augmentation. In *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 68–82. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Taha Zerrouki. 2010. pyarabic, an Arabic language library for Python.